# ЛАБОРАТОРНАЯ РАБОТА №3

## Анализ и прогноз развития рынка жилой недвижимости с использованием возможностей ПО Statistica 12.0.

**Цель работы** – знакомство с модулем *Временные ряды и прогнозирование (Time Series/Foregasting)*. Подбор модели авторегрессии и скользящего среднего к заданному временному ряду. Прогноз по полученному уравнению и оценка адекватности модели.

**Справочные сведения.** Изучение временного ряда на практике чаще всего имеет своей целью подбор статистической модели, описывающей временной ряд, и предсказание будущих его значений. Методы анализа временных рядов широко представлены во многих универсальных статистических пакетах STADIA, STATGRAPHICS, SPSS, STATISTICA. Но анализ временных рядов – это очень специфическая область статистики, отличающаяся по кругу задач и методов их решения, а также по составу пользователей, применяющих эти методы.

Одним из методов анализа временных рядов являются модели авторегрессии и скользящего среднего, которые оказываются особенно полезными для описания и прогнозирования процессов, проявляющих однородные колебания вокруг среднего значения. Однако эта модель подходит только для стационарных рядов, среднее, дисперсия и автокорреляция у которых примерно **постоянны во времени.**

В большинстве временных рядов члены ряда зависят друг от друга. В гидрологических рядах значимые внутрирядные связи наблюдаются чаще у соседних членов и быстро уменьшаются с увеличением расстояния между ними. На этом свойстве влияния предыдущего состояния процесса на будущее базируются модели авторегрессии. Математически это свойство можно выразить уравнением

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_p + \varepsilon_t,$$

где $y_t$ – значение y в момент времени t; $\varphi_i$ – коэффициенты уравнения (i = 1,2,…,p); p – порядок авторегрессии; $\varepsilon_t$ – случайная величина.

В модели скользящего среднего в отличие от предыдущего способа предполагают, что каждый элемент ряда подвержен суммарному воздействию случайных предыдущих ошибок $\varepsilon_i$:

$$y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

где $y_t$ – значение y в момент времени t; $\theta_i$ – коэффициенты уравнения (i = 1,2,…,q); q – порядок модели скользящего среднего; $\varepsilon_t$ – случайная величина.

Объединённая модель авторегрессии и интегрированного скользящего среднего (ARIMA) была предложена Боксом и Дженкинсоном в 1976 г. Аббревиатура ARIMA, образованная от слова autoregression (AR) и английского названия moving average (MA) для скользящего среднего, стандартно используются в литературе и статистических пакетах.

В русском переводе название модели – АРПСС, что также является аббревиатурой: АР – авторегрессия, СС – скользящее среднее.

В пакете Statistica анализ и прогноз по модели авторегрессии и скользящего среднего осуществляется в модуле *Временные ряды и прогнозирование (Time Series/Foregasting)*. В процессе работы генерируется большое количество диалоговых и вспомогательных окон, таблиц, графиков. Для успешной навигации по ним приводится схема переходов между четырьмя основными диалоговыми окнами первого уровня (рис. 3.1).

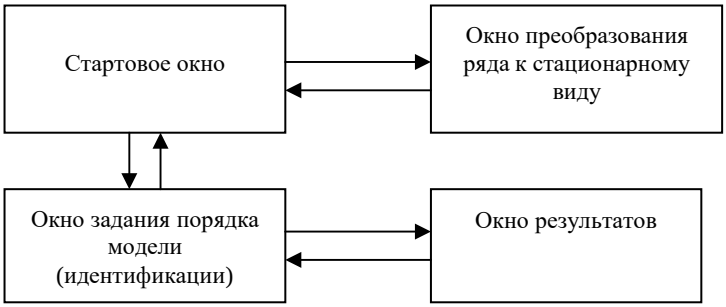


Рис. 3.1. Схема переходов между диалоговыми окнами

**АРПСС (ARIMA)** – сложная модель, требующая определённого опыта в использовании. Анализ и прогнозирование поведения временного ряда состоит из нескольких этапов, включающих выполнение различных статистических и сервисных процедур:

1. Восстановление пропущенных данных.
2. Преобразование ряда к стационарному виду.
3. Идентификация модели, т.е. подбор порядка модели p и q.
4. Оценка параметров модели.
5. Проверка адекватности модели.
6. Прогноз по модели.

**Последовательность выполнения работы:**

1. Загрузите пакет Statistica. Установите удобный Вам режим сохранения результатов работы с помощью Диспетчера вывода *Файл → Диспетчера вывода (File → Output Manager).*

2. Создайте новую Рабочую книгу с таблицей для исходных данных *Файл → Создать → Создать новый документ* (*File → New → Create New Document*) и скопируйте в неё временной ряд с данными гидрометеорологических наблюдений в соответствии с номером Вашего варианта. Рекомендуется иметь, как минимум, 50 наблюдений в файле исходных данных.

Созданную рабочую книгу лучше сразу переименовать и сохранить в своей папке. Тогда в течение сеанса ее имя будут располагаться в верхней части списка кнопки *Добавить в рабочую книгу (Add to Workbook),* которую Вы будете использовать для добавления результатов в рабочую книгу.

3. Каждому ряду дайте уникальное имя вместо стандартных Var1, Var2…, а в поле *Длинная метка или формула (Long Name)* Вы можете занести любую полезную информацию о ряде, например, период наблюдений, характеристики поста и т.д. Переименование осуществляется в окне спецификаций переменной, которое можно вызвать двойным щелчком ПК мыши на имени столбца.

Вызовите модуль для работы с временными рядами *Анализ → Углубленные методы анализа → Временные ряды и прогнозирование (Statistics → Advanced Linear/Nonlinear Models → Time Series/Foregasting).*

4. Восстановление пропущенных данных. Замените временно десятое по счету наблюдение на константу отсутствия информации –9999. Все пустые ячейки пакета по умолчанию содержат этот код. Восстановите отсутствующее данное пятью различными способами, которые предлагаются в стартовом окне модуля на вкладке *Обработка пропусков (Missing data)*, предназначенном для обработки пропущенных значений
(рис. 2):

- общее среднее (среднее по всем наблюдениям);
- интерполяция по ближайшим точкам (двум соседним);
- среднее, рассчитанное по N ближайшим точкам (N задаете сами);
- медиана ряда, составленного из N ближайших точек;
- восстановление по уравнению линейной регрессии.

Восстановленное значение можно посмотреть в специальной таблице. Для этого надо перейти на другое диалоговое окно (рис. 3.2) кнопкой OK и на вкладке *Графики (Review & Plots)* заказать *Просмотр выделенной переменной (Review highlighted variable).*

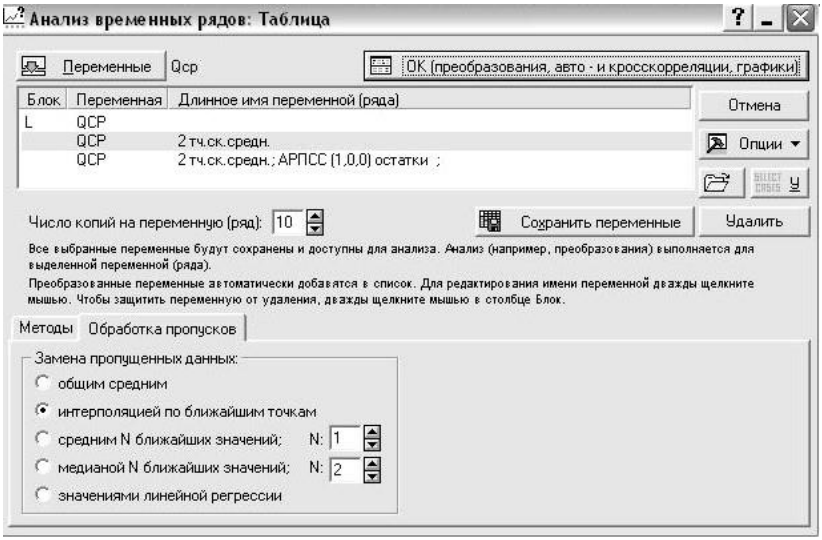


Рис. 3.2. Стартовое окно модуля. Меню для восстановления пропущенных данных

Обратите внимание! Если у Вас в списке несколько рядов, то выбранная Вами команда просмотра будет относиться к выделенному ряду.

Выпишите значения, восстановленные разными способами, сравните их с исходными и дайте заключение о точности восстановления.

5. Преобразование ряда к стационарному виду. Модель АРПСС (ARIMA) применима только к стационарным рядам, поэтому сначала надо построить график для своих данных и, если ряд не отвечает этому требованию, то многократными преобразованиями привести его к стационарному виду.

Для приведения ряда к стационарному виду перейдите в окно для преобразований кнопкой *OK (преобразования, авто- и кросскорреляции) (transformations,.....)* и на вкладке *Графики (Review & Plots)* выберите команду для построения график исходного ряда. Там же поставьте галочку у режима строить *График после каждого преобразования (Plot Variables after each transformation).*

Теперь результат каждого преобразования будет автоматически высвечиваться на экране в виде графика, и Вы сможете визуально оценивать стационарность ряда (рис. 3.3).

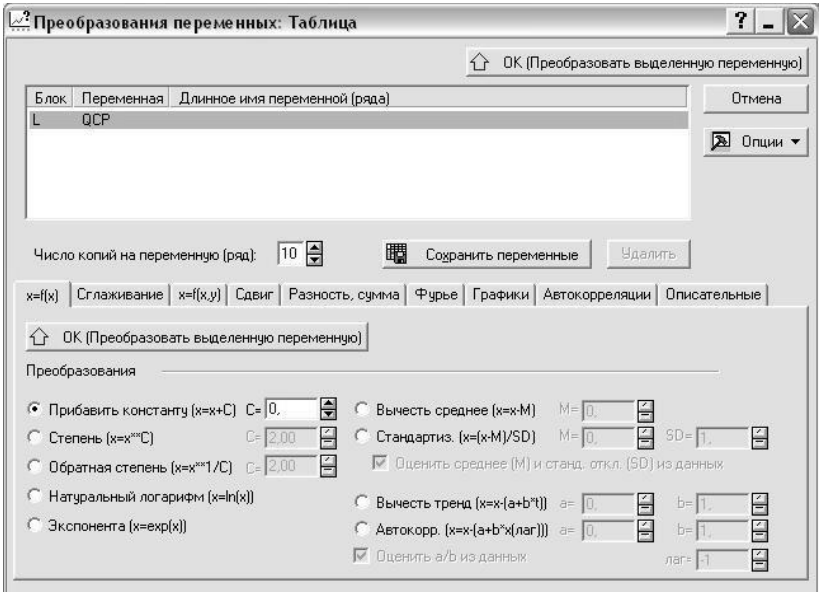


Рис. 3.3. Диалоговое окно для преобразования ряда

Рекомендуются следующие преобразования:

• логарифмирование для стабилизации дисперсии и уменьшения амплитуд колебания временного ряда (вкладка x = f(x));

• удаление тренда (вкладка x = f(x));

• взятие разности первого или более высокого порядков на вкладке *Разность (Difference)* для удаления линейного тренда;

• сглаживание (вкладка *Smoothing*);

• вычитание среднего, стандартизация ряда (вкладка x = f(x)).

Для проверки стационарности полезно также строить для каждого преобразованного ряда график автокорреляционной функции (вкладка *Автокорреляции – Autocorrs*). У стационарного ряда график автокорреляционной функции стремиться к нулю с увеличением лага.

Не забывайте основное правило всех диалоговых окон модуля:

6. Идентификация модели. Следующим шагом является подбор порядка модели АРПСС (ARIMA). Модель – это уравнение, по которому будут

вычисляться прогнозируемые величины. Идентификация модели – это выбор конкретного вида уравнения. Вы должны указать программе, сколько слагаемых будет в уравнении авторегрессии (p), а сколько – в уравнении скользящего среднего (q). Величины p и q называются параметрами модели. Например, при p = 1 и q = 1 уравнение модели будет выглядеть следующим образом

$$y_t = \varphi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \text{const},$$

где $y_t$ – значение исследуемой величины в момент времени t; $y_{t-1}$ – значение исследуемой величины в предыдущий момент времени; $\varepsilon_{t-1}$ – значение случайной составляющей исследуемой величины в момент времени t–1; $\varphi_1$, $\theta_1$ – коэффициенты уравнения, которые будут подбираться по методу наименьших квадратов.

Вернитесь в стартовое окно и на вкладке *Методы* щелчком на кнопке *АРПСС и автокорреляционные функции (ARIMA & autocorrelations)* перейдите в диалоговое окно, предназначенное для идентификации модели (рис. 3.4).

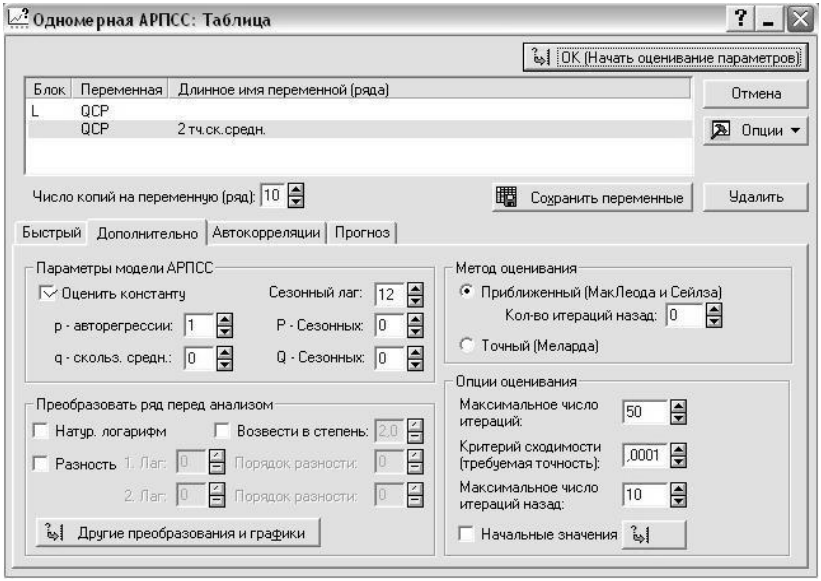


Рис. 3.4. Диалоговое окно для задания порядка модели

Закажите здесь количество членов в уравнении модели, например, p = 1, q = 0 и наличие свободного члена (галочка в поле *Оценить константу – Estimate constant*). В дальнейшем, если модель получится неудачной, будете варьировать значения p и q с учетом рекомендаций по виду графиков автокорреляционной функции (АКФ) и частной автокорреляционной функции (ЧАКФ).

График автокорреляционной функции называют также коррелограммой (рис. 3.5). На этом графике кроме самой АКФ, значения которой изображаются в виде столбиков, указаны доверительные интервалы для коэффициентов автокорреляции, которые в пределах доверительной трубки незначимо отличаются от нуля.

*Рекомендации по подбору порядка модели на основе анализа графиков АКФ и ЧАКФ.* Предварительный вывод о порядке модели скользящего среднего можно сделать по числу первых q значимых значений автокорреляционной функции (рис. 3.5). Указанное правило хорошо, если подобранный порядок невелик, например, от одного до трёх-четырёх.


Рис. 3.5. Графики автокорреляционной и частной автокорреляционной функций

Для подбора порядка авторегрессии p большую информацию даёт вид частной автокорреляционной функции. Если значим первый коэффициент автокорреляции, то p = 1. Если значимы два первых коэффициента, то p = 2.

Решение, какие значения задать для p и q, является не простым и требует эксперимента с различными моделями. Тем не менее, можно дать следующие практические рекомендации, обобщающие оба графика:

- Задать параметр p = 1, если АКФ экспоненциально убывает, а ЧАКФ имеет резко выделяющееся значение на лаге 1, нет корреляции на других лагах.
- Задать параметр p = 2, если АКФ имеет форму синусоиды или экспоненциально убывает. ЧАКФ имеет резко выделяющиеся значения на лагах 1 и 2.
- Задать параметр q = 1, если АКФ имеет резко выделяющееся значение на лаге 1, нет корреляции на других лагах. ЧАКФ экспоненциально убывает.
- Задать параметр q = 2, если АКФ имеет резко выделяющееся значение на лаге 1 и 2, нет корреляции на других. ЧАКФ имеет форму синусоиды или экспоненциально убывает.
- Задать p = 1 и q = 1, если АКФ экспоненциально убывает с лага 1, ЧАКФ экспоненциально убывает с лага 1.

7. Оценивание параметров модели. Теперь, когда Вы задали параметры модели и таким образом определили количество неизвестных коэффициентов в уравнении, можно запустить процедуру их нахождения (оценивания) кнопкой *Ok (Begin parameters estimation).*

Метод оценивания выбирается в левой нижней части окна (см. рис. 3.4). Система предлагает две вычислительные процедуры, реализующие метод максимального правдоподобия, приближенную и точную (*Exact*).

Если итерационный процесс вычисления коэффициентов сошелся, то появится окно с результатами вычислений *Результаты АРПСС (Single Series ARIMA Results)*. В информационной области диалогового окна результатов (рис. 6) высвечиваются следующие сведения:

- имя ряда наблюдений;
- перечень преобразований *(Transformations)*, которым подвергался ряд;
- вид модели: Model (p, d, q), где d – число преобразований типа взятия разностей первого или более высоких порядков;
- количество наблюдений в исходном ряду *(No. of obs)*;
- начальное и конечное значения остаточной суммы квадратов (SS) и средний квадрат остатков (MS);
- числовые значения коэффициентов уравнения и их стандартные ошибки.

Коэффициенты уравнения модели устойчивы, если они, по меньшей мере, в два раза превышают свои стандартные ошибки. Красным цветом выделяются значимые коэффициенты.

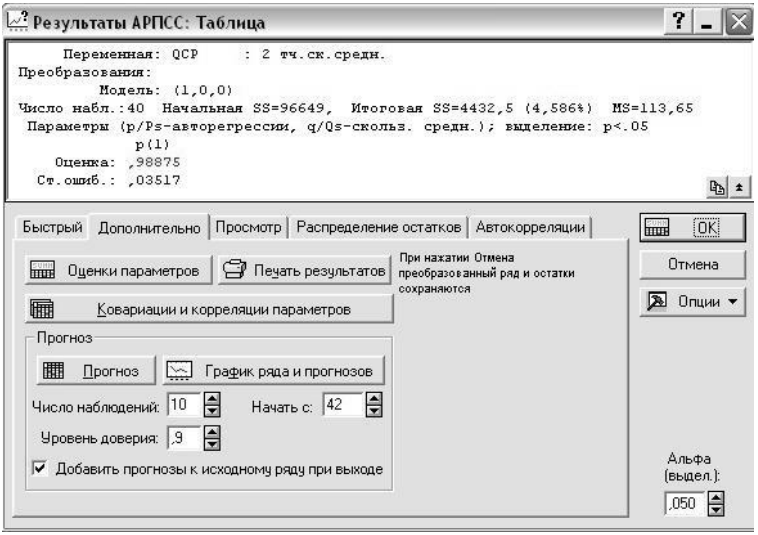


Рис.3.6. Окно результатов, заказа прогноза и анализа остатков

8. Прогноз по модели. Параметры прогноза можно задать в окне результатов (рис. 3.6) на вкладке *Дополнительно (Advanced)* в поле *Прогноз (Foregasting)*:

- заблаговременность – *Число наблюдений* (*Number of Cases*);
- номер элемента ряда, с которого предполагается начать прогноз (*Start at Case*);
- доверительную вероятность прогноза – *Уровень доверия (Confidence level).*

График ряда с добавленными спрогнозированными значениями красного цвета и доверительными интервалами для них (рис. 3.7) можно получить, щелкнув на кнопке *График ряда и прогнозов (Plot series & forecasts)*. Рекомендуется построить несколько моделей и выбрать лучший вариант.

Существуют два варианта задания ряда для прогноза, связанные с ограничениями модели. Если для прогноза задаётся имя исходного ряда, то модель автоматически учитывает преобразования только типа взятия разностей (параметр d) и выдает результаты для исходного ряда.

Если при приведении ряда к стационарному виду использовались преобразования других типов, то прогноз надо заказывать для преобразованного ряда. Так как результаты промежуточных преобразований хранятся в модуле, то

пересчет спрогнозированных величин на ряд наблюдений не представляет сложностей.

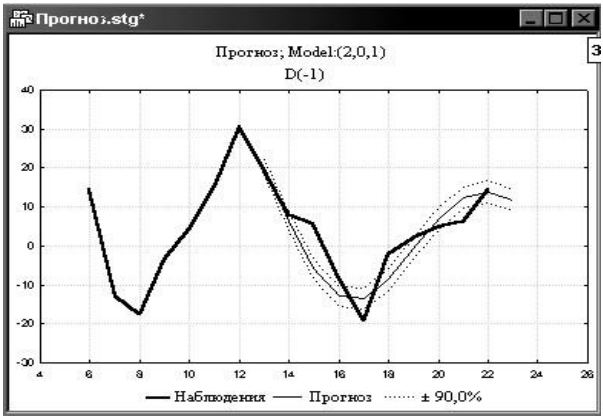


Рис. 3.7. Пример прогноза по модели с порядком p = 2 и q = 1

9. Анализ адекватности модели данным. К сожалению, единого общего правила для этого анализа нет. Более или менее обоснованное решение можно принять, сравнив имеющиеся наблюдения со значениями, полученными с помощью подобранной модели.

Разности между наблюдаемыми и предсказанными значениями называют **остатками**.

Анализ остатков позволяет получить представление, насколько хорошо подобрана сама модель и насколько правильно выбран метод оценки коэффициентов.

Предполагается, что модель адекватна, если выполняются 2 требования: 1) остатки независимы,

2) остатки распределены по нормальному закону.

Для проверки независимости остатков обычно используют критерий серий, критерий Дарбина-Уотсона, автокорреляционную функцию. В модели ARIMA для этих целей предлагается использовать автокорреляционную функцию (коррелограмму).

Для проверки нормальности распределения остатков чаще всего используется график на нормальной вероятностной бумаге, а также критерии Колмогорова-Смирнова, хи-квадрат и т.д.

В окне результатов закажите график остатков на вкладке *Просмотр (Review & residuals)*, гистограмму ряда остатков и график на нормальной вероятностной бумаге (вкладка *Распределение остатков – Distribution of Residuals*). Затем постройте график автокорреляционной функции остатков (вкладка *Автокорреляции – Autocorrelations*). Если коэффициенты автокорреляции незначимы (не выходят за пределы доверительного коридора) и расположены хаотично, то остатки независимы.

Если модель не адекватна, то придется начать моделирование сначала, с подбора новых p и q или с более раннего этапа – преобразования ряда к стационарному виду.

**Отчетные материалы:**

1. Теоретическая часть по моделям авторегрессии и скользящего среднего.
2. Анализ результатов восстановления пропущенных данных.
3. Уравнение подобранной модели.
4. Графики с прогнозом.
5. Графики, подтверждающие адекватность модели: график и гистограмма остатков, нормальный вероятностный график (*Normal Probability Plot*), коррелограмма остатков.

## ЛИТЕРАТУРА


1. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере / под ред. В.Э. Фигурнова. М. : ИНФРА, 1998. 528 с.
2. *Шелутко В.А.* Численные методы в гидрологии. Л. : Гидрометеоиздат, 1991. 238 с.
3. *Берестнева О.Г., Муратова Е.А., Уразаев А.М.* Компьютерный анализ данных. Томск : Изд-во ТПУ, 2003. 204 с.
4. *Боровиков В.П., Боровиков И.П.* Statistica.Статистический анализ и обработка данных в среде Windows. М. : Филинъ, 1997. 608 с.
5. *Рождественский А.В., Чеботарев А.И.* Статистические методы в гидрологии. Л. : Гидрометеоиздат, 1974. 424 с.